DOCUMENT RESUME

ED 174 682 TM 009 654

AUTHOR Hills, John R.
TITLE Using Empirical Data to Set Cutoff Scores.
PUB DATE Apr 79
NOTE 12p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (San Francisco, California, April 9-11, 1979)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Academic Standards; Criterion Referenced Tests; *Cutting Scores; *Mastery Tests; Probability; Scores; Student Testing
IDENTIFIERS Aptitude Treatment Interaction; *Test Length

ABSTRACT

Six experimental approaches to the problems of setting cutoff scores and choosing proper test length are briefly mentioned. Most of these methods share the premise that a test is a random sample of items, from a domain associated with a carefully specified objective. Each item is independent and is scored zero or one, with no provision for guessing. Developers and starting points of these methods are: (1) Berk--instructed and uninstructed student groups; (2) Millman--true scores; (3) Hambleton and Novick--type of student placement error; (4) Phaner, modified by Wilcox--indifference zone; (5) Davis and Diamond--Bayes's Theorem; and (6) Visco--observed scores. Unlike the other methods, Visco considers finiteness of domain, guessing, and setting cutting scores for nonmastery. Visco raises the issue of emphasis on decisions about mastery to advance students, versus certain decisions about nonmastery so that remediation can be provided. There are many complex issues to consider before teachers can use cutting scores effectively. (CP)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED174682

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

USING EMPIRICAL DATA TO SET CUTOFF SCORES

Abstract

This paper is part of a symposium on setting cutoff scores. One speaker agreed to address the use of judgments to set standards; another agreed to discuss Bayesian methods which use both judgments and data; and the third agreed to present methods based just on empirical data. This is the third paper.

Six different methods are presented, noting their common assumptions and their different approaches, as well as the similarities among their conclusions. The current status of use of ATI's is discussed, and it is noted that these methods are all too complex for use by the classroom teacher who probably sets more cutoff scores than any other group or agency in our culture.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

John R. Hills

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM009 654

USING EMPIRICAL DATA TO SET CUTOFF SCORES

by

John R. Hills

Florida State University

The 1974 edition of Standards for Educational and Psychological Tests published by this organization, the American Educational Research Association, and the American Psychological Association takes a clear stand on the setting of cutoff scores. It states:

"I.4. If specific cutting scores are to be used as a basis for decisions, a test user should have a rationale, justification, or explanation of the cutting scores adopted. Essential." (p. 66)

It goes on to indicate that many kinds of arguments might be used. In a content-referenced interpretation of a mastery test, the Standards suggest that one might consider the probability that a confidence interval around the obtained score includes a perfect score.

If interpretations are referenced against an external criterion, the Standards suggest that the cutting score might be based on a designated probability of achieving a specified level of success.

If discrimination between groups is desired, the Standards suggest that decision theory be used to maximize the discrimination.

In a selection situation, they suggest that cutting scores might be set in terms of prediction of yield.

The intent of the Standards is "to recommend that test users avoid the practice of designating purely arbitrary cutting scores they can neither explain nor defend."

Paper presented to the Annual Meeting of the National Council on Measurement in Education, San Francisco, April, 1979.

The suggestions from the Standards are not very different from the discussion in the ACE's second edition of Educational Measurement. But that presentation was written a decade ago. There have been some developments since then. I will review some of them today.

Before doing that, let me consider briefly the current status of the decision theory model for setting cutting scores. As far as I know, very little application of that kind of thinking has occurred in school settings. The idea was basically that aptitude-treatment-interactions were to be investigated and used to set the cutting scores.

The recent work of Cronbach and Snow entitled Aptitudes and Instructional Methods devotes over 500 pages to research related to instructional ATI's. Listen to a few of their comments.

Investigators should be looking for at least interactions of the fourth order. They are seeking something like this:

> "With subject matter of this nature,
> inductive experience of this type,
> in this amount,
> produces this pattern of responses,
> in pupils at this level of development."
> (page 517, from Cronbach, in Schulman & Keislar, 1966, p. 77)

The interaction is among such things as ability, age, subject matter, treatment, and outcome.

Here is another of their comments:

> "No Aptitude x Treatment interactions are so well confirmed that they can be used directly as guides to instruction." (p. 492)

And here is another warning: The crossing of regression lines for classes assigned to different treatments tells us how we might assign classes, but not how we should assign individuals. Using a rule to place individuals will assemble classes in a new way; the study we used to set the cutting score tells us nothing about ATI of the newly assembled classes. (See about page 105.)

Calling attention to these problems is not meant to discourage test users from thinking of placement problems in terms of the ATI model. It still makes more sense than less sophisticated thought patterns. However, our initial impression that headway would be rapid with this new understanding did not appreciate just how complicated the placement problem is.

During the last decade or so, a number of interesting approaches to setting cutting scores have been suggested and explored. Berk (1976) suggested an empirical method of setting cutting scores that is based on the assumption that students who have been instructed are masters and students who have not been instructed are nonmasters. (I wish I could always safely assume the former about my students.) Berk suggests that we try different cutting scores on a test to see which score results in the most accurate sorting into instructed and uninstructed groups. The technique appears to be the same as setting a cutting score on the basis of performance in a subsequent course. This idea was discussed in the second edition of Educational Measurement back in 1971. Berk's criterion is just different and novel.

Often the problem of setting the cutting score is tied closely to the problem of choosing a proper test length. After all, if the cutting score has to be higher than the test length, we either have a great solution or a very difficult problem.

Most of the new ideas on setting cutting scores are based on domain sampling. They seem to have this set of premises in common:

1. We have a domain of items associated with a carefully specified objective, or other structure, which either allows us to catalog the items or else allows us to decide clearly whether any new item belongs to the domain.

2. Every item in the domain is independent of the others. Being presented with one does not influence performance on any other item.

3. A test is a random sample of items from the domain.

4. Items are scored zero or one; most of the methods do not take the possibility of guessing into consideration.

These assumptions permit use of a model in which the student's command of a domain is like an urn. His responses to items are like balls drawn from the urn. There are two colors of balls. The problem is to decide what proportion of balls in the urn are of a particular color. What is known is how many balls are drawn and how many among those drawn are of the desired color.

Within that frame of reference, different scholars attack the problem in different ways. Millman starts with the true score. He considers the problem of determining how long a test must be to keep errors of misclassification to desired levels for students at different levels of true score. The length of the test and percent of misclassifications depend on how many items the student is allowed to get wrong.

For example, suppose you want to know whether a student could answer correctly 80% of the items in the domain. Millmar's tables tell you such things as that if you are willing to be wrong about 25% of the time for students who really know 70% or 90% of the items, you can give 8 randomly chosen items and decide that anyone who gets 7 or 8 correct is a master. If you are willing to give a much longer test, you can detect with fewer errors people who are closer to the 80% value but not exactly at it. For

instance, with a set of 100 items, setting a cutting score at 80 correct gives 15% errors for people who really know only 75% and 7% errors for people who really know 85%. To set the cutoff score, then, you need to decide what kind of errors, of what degree, you can tolerate in what amounts. The cutting score, of course, depends on the number of items in the test, also.

Hambleton and Novick (1973) suggested that one consider not only whether an error is made in classifying a student, but which kind of error is more serious. They suggest that one should consider the ratio of loss for a false positive to loss for a false negative. Huynh (1976) described how to do this. Setting the cutoff score requires determining the loss for each possible score and choosing as a cutoff score the point where the loss is the least.

Fhaner (1974), modified by Wilcox (1976), suggests that the cutting score be set in terms of an indifference zone, the idea is that for students whose score is near the cutoff, there is not much loss if they are classed one way or the other. So we can be indifferent as to whether those students are misclassified. For students far from the cutoff, we want to have little danger of making errors.

In Wilcox's development, we must set the desired probability of making a correct decision. However, if you wish, you may specify different probabilities for being correct about false positives than about false negatives, which is the same idea as having different losses for the two kinds of errors.

Using Wilcox's procedure we get results such as the following:

If we want to be sure that a student knows the answers to 80% of the items in the domain, and if we want to have a probability of .75 of

correctly classifying any student who could answer 70% or less and any student who could answer 90% or more, then we need a test of 8 randomly selected items, and we class as masters those who get 7 and 8 correct. We classify as nonmasters those who get 6 or fewer correct. Notice that with this method, we have a new parameter to set, the size of our indifference zone.

Davis and Diamond (1974) use Bayes's Theorem to calculate the lowest acceptable cutting score on a test of a given length if the lowest acceptable competence level is specified and the desired proportion of correct classifications is specified.

For example, if the minimum level of competence we are willing to accept is knowing the correct answer to 80% of the items in a domain, and if we want to categorize the pupils correctly 85% of the time, then we need to use a test with 12 items, and the examinee must get all 12 correct to be called a master. If we set competence at knowing the correct answer to 90% of the domain, and if we want to be correct 85% of the time in our decisions about students, then for tests up to 20 items in length the examinee must get a perfect score to be classed a master.

Most of the suggestions we have considered so far have acted as though one cutting score was to be used to class as a master vs. a nonmaster. It is conceivable that one could set a cutting score in terms of probabilities of correct decisions about masters, without trying to decide whether those not classed as masters were truly nonmasters. This would be a little like Wilcox's approach with an indifference zone. Visco, in a 1977 dissertation at Florida State University, took that approach. Unlike Millman, he started from the observed score, and unlike Davis and Diamond, he chose not to use Bayes' theorem.

He started with the observed score and a specified confidence interval and solved for the lowest true score that would be in the interval. For example, he used the binomial distribution to determine that if a student answers four items out of five from an infinite sized domain, one is 95% confident that his true proportion correct on the entire domain is at least .42. If one wanted to be 95% confident that the true score correct on the entire domain was .8, he would have to use a much longer test, and the student would have to get a high proportion of the items on that longer test correct. The cutting score in this method depends not only on test length, degree of confidence desired, and level of criterion, but also on whether the domain is finite or infinite and whether guessing is possible.

In addition, Visco then turns the problem around and considers setting cutting scores for nonmastery. In such a case, one solves for the upper limit of the confidence interval instead of the lower limit. He concludes that if a score is low enough, one can have a reasonable degree of confidence that the student is correctly classed as a nonmaster. It would seem that when students are placed in remediation, this is really the decision in which we are most interested. Most of the effort in research concerning setting of cutting scores has been stimulated by instructional design problems in which the issue was whether to advance a student to the next module. Thus mastery got the attention. It may well be that in the real world of schools far more decisions are made about putting nonmasters into remedial work than are made using tests designed to determine if a student is ready for an advanced class or module. Thus Visco's direction of our attention to setting cutting scores for nonmastery may be very beneficial.

Visco gives tables for interesting values. For example, if our criterion for mastery is set at 80% correct, and if we want to be at least 90% confident, then a student would have to get all eleven items correct on an eleven-item test from a domain of infinite size. On the other hand, if we set 60% correct as our criterion for nonmastery, and we want to be 90% confident that a student's true score on the entire domain would be less than 60% before we subject him to remediation, then we can use a three-item test if he cannot get any of those items correct. If he can only get one correct our of five items, we can be 90% confident of nonmastery. And so on.

Now, if the level of the criterion for nonmastery is set high, we can use relatively few items to declare with confidence that low scorers are nonmasters. If the criterion level for mastery is set low, we can use relatively few items to declare with confidence that high scorers are masters. But in the interesting situation in which we want to set a low criterion for nonmastery, we will need a lot of items to be confident even for those who score very low. Likewise, if we want a high criterion for mastery, even high scorers will have to demonstrate their proficiency on many items in order for us to be confident of their mastery. The most interesting situations require many items, and those who get about half the items correct are not easily classified with confidence.

Visco's approach requires us to think about some new issues. Are we really more interested in deciding mastery in order to advance students, or are we more interested in deciding about nonmastery with confidence so that we can soundly apply remediation. Once we decide that, we have to decide where we want the criterion. For nonmastery, do we want a low criterion score? If so, we'll need a lot of items in order to be confident.

If we are willing to set a high criterion score even though we are interested in nonmastery, we can get along with fewer items.

In looking back over these various methods, it is interesting that none of them offers support to the notion that in criterion-referenced testing, we need only a few items to decide whether a student has mastered an objective. Perhaps the Bayesian procedures that assume some prior knowledge will move in that direction, but I don't recall hearing the instructional designers using Bayesian priors to justify the suggestions they make and the limited numbers of items they use for criterion-referenced tests.

In closing, there is one final perplexing thought. The most widespread users of tests with cutting scores are surely the classroom teachers who make up their own tests or use the tests of publishers of instructional systems. How are they going to set soundly the cutting scores on their own tests, or evaluate the soundness of the cutting scores given them by publishers? There are so many complex issues to consider, and those teachers are already so overloaded with large classes, and even with extra duties such as supervising playgrounds, lunchrooms, and football games. But if we don't have a way to tell them how to make these kinds of decisions effectively, what are the alternatives? Can we tolerate them?

REFERENCES

Berk, R. A. "Determination of optimal cutting scores in criterion-referenced measurement." Journal of Experimental Education, 1976, 45, 4-9.

Cronbach, L. J. & Snow, R. E. Aptitudes and Instructional Methods. New York: Irvington, 1977.

Davis, F. B. & Diamond, J. J. "The preparation of criterion-referenced tests." In Harris, et al. (Eds.) Problems in Criterion-Referenced Measurement. Los Angeles: University of California, 1974, 116-138.

Fhaner, S. "Item sampling and decision-making in achievement testing." British Journal of Mathematical and Statistical Psychology, 1974, 27, 172-5.

Hambleton, R. K. & Novick, M. R. "Toward an integration of theory and method for criterion-referenced tests." Journal of Educational Measurement, 1973, 10, 159-170.

Huynh, H. "Statistical consideration of mastery scores." Psychometrika, 1976, 41, 65-78.

Joint Committee of APA, AERA, NCME. Standards for Educational and Psychological Tests. Washington: American Psychological Association, 1974.

Millman, J. "Passing scores and test lengths for domain-referenced measures." Review of Educational Research, 1973, 43, 205-216.

Visco, R. J. The Classification of Students with Respect to Achievement, with Implications for State-Wide Assessment. Unpublished Dissertation, Florida State University, Tallahassee, Florida, 1977.

Wilcox, R. A. "A note on the length and passing score of a mastery test." Journal of Educational Statistics, 1976, 1, 359-364.